Design of a Motility Assay: The Statistical Point of View

Part 2: Statistical Data Analysis

Introduction:

This is the second part of a three-part tutorial on the design and subsequent analysis of experiments such as motility assays.
In the first part we generated some synthetic data that represented the movement of a bacterium like b-sub. In the second part we look into the basics of Bayesian data analysis and also on the ideas of quality control and design of experiment. To simplify things we will assume that the data were obtained with an extraordinary level of precision.
In the last tutorial, we will study the more complex – and more realistic - case where the data acquisition process only gives us imperfect access to the data. As you will see the quality of the predictions that we can make is degraded and from an experimental point of view we need to gather more data.

Pre-Analysis Data Processing

The outcome to your motility assay is a movie – i.e a sequence of frames – onto which a bacterium-tracking algorithm can be applied.
For the moment we will not question the output to the tracking algorithm. The times and positions it yields are therefore considered exact.

[Discuss this statement]

In order to analyse the data, they must be broken into relevant, manageable subsets

- Design a routine converting the trajectory of a bacterium into the relevant times, angle, velocities etc…
- Apply your routine to your synthetic dataset(s)
- ***Crude Error Checking***: Plot the relevant histograms and compare them to histograms of the synthetic dataset(s) generated in **Tutorial 1** (they should be the same)

With the kind of motion mentioned in **Tutorial 1**, the trajectory can be split into four sets of independent data. However, we can expect velocity and time to be linked.

- **[Discuss this statement]**
- In the likely event they are indeed linked, modify the pre-analysis data processing – and the error checking – accordingly.

Standard Parameter Estimation Methods

The Usual Method:

The parameters of a statistical distribution can be estimated from a set of samples in many different ways. The most common way is to apply what is called an estimator to the distribution in order to extract its parameters. Probably the simplest family of methods derive from the principle of Maximum Likelihood (**ML**)

- What is the principle of Maximum Likelihood?
- In general, how do you compute the **ML** estimator of the vector parameter α of a distribution P given the available data $D = (D_1,\ldots,D_n)$?
- What are the **ML** estimators of standard distributions such as
    - The Gaussian Distribution
    - The Exponential Distribution
    - The Poisson Distribution
    - The Uniform Distribution
    - The Von Mises Distribution
    - The Ricean Distribution
    - The Maxwellian Distribution

Unfortunately we never have access to the entire distributions (that is an infinite amount of data representative of the distribution), only to a few data representative of the distribution. The amount of data we use to estimate the parameters is therefore a crucial factor in the accuracy of the outcome.

- Apply the relevant estimators to your data with various amounts of data
- For instance use 10 samples, 50 samples, 100 samples
- Does the order of your data have a significant influence?
- Apply an estimator to the wrong distribution – make the amount of data vary
- **Discuss the pros and cons of the ML Approach**

A Possible Alternative:

Another possible method is to recombine the moments of the distribution in such a way that the outcome is a parameter of the distribution.

- What is the definition of m_n the n^{th} moment of a distribution?
- Often we prefer centred moment for n>1. What is their definition?
- How do we call the centred moments of second, third and fourth order?
- What do they measure?

Recombining the moments $m_1 \ldots m_n$ in order to estimate the parameters $\alpha_1 \ldots \alpha_p$ consists in identifying the functions $g_1 \ldots g_p$ such as

$$\begin{cases} m_1 = f_1(\alpha = (\alpha_1,....,\alpha_p)) \\ \quad . \\ \quad . \\ \quad . \\ m_n = f_n(\alpha = (\alpha_1,....,\alpha_p)) \end{cases} \Rightarrow \begin{cases} \alpha_1 = g_1(m_1,....,m_n) \\ \quad . \\ \quad . \\ \quad . \\ \alpha_p = g_p(m_1,....,m_n) \end{cases}$$

- Apply the method to the Gaussian, Exponential, Poisson distributions…
- The functions $g_1 \ldots g_p$ are not unique - see the case of the exponential distribution. How do you think the discrepancies between the various formulae can be exploited?

Again we never have access to the entire distributions, only to a few data that we think are representative of the distribution. The amount of data we use to estimate the parameters is again crucial for the accuracy of the outcome.

- Calculate the first four moments with various amounts of data
- For instance use 10 samples, 50 samples, 100 samples
- Does the order of your data have a significant influence?
- Apply the recombination to the computed sample-moments of distributions of your choice
- Compare the accuracy of the approach to the results obtained with **ML**.
- **Discuss the pros and cons of the method**

The Bayesian Approach

The Bayesian Philosophy

The other way of looking at data is the Bayesian approach. It is still a controversial interpretation of the concept of probability –if you read the chapter I have posted, you will see that in my opinion it is the only consistent, honest interpretation.
A Bayesian probability is commonly seen as a 'degree of belief' and a distribution is seen as a summary of what we know about a phenomenon. At the heart of the theory lies a theorem (Bayes' theorem) and an update rule.

- What is Bayes' theorem?
- What is the meaning of the terms involved in the theorem?
- What is Bayes' update rule? What is its meaning?

An Example of Bayesian Parameter Estimation

One of the easiest applications of Bayesian analysis is the estimation of the parameters of a statistical distribution. To see how it is done let us deal with a simple example: the exponential distribution

- What is the exponential distribution of parameter λ?
- Write the general expression for likelihood of the sample vector $X=(x_1,\ldots,x_N)$
- It is customary to assign to scale parameters such as λ a prior distribution $P(\lambda)=1/\lambda$. Write the posterior of λ
- Find the general expression of the maximum of the posterior
- **Bonus Question:** $P(\lambda)=1/\lambda$ is not normalised… What can be done about it?
- Draw 10 samples $x_1,\ldots,x_{10}$ from the exponential distribution $\lambda=1$
- For these 10 samples plot the posterior of λ
- Calculate its maximum, its expected value and its standard deviation

Influence of the Quantity of Data and the Hypothesis

- Re-apply the analysis with 50 samples, 100 samples.
- What do you observe?
- I want to estimate λ with a 5% error: how many samples do I need?
- Now, Draw 10, 50, 100 samples with a Gaussian distribution of mean 1 and standard deviation 1.
- Apply the previous Bayesian analysis to these samples
- What do you observe?

Interlude: The Bayesian Philosophy Pt 2

As you must surely have realised there are several ways to estimate the parameters from their posterior distribution. This may seem confusing but there are very good reasons for this.
The first reason is that although the Bayesian approach is closely linked to decision theory, it is not solely concerned with reaching decisions. The main purpose is the construction of the posterior distributions –the summaries of what we know. The posteriors will be updated upon arrival of new data. There is therefore no point in reaching a decision – which inevitably leads to a loss of knowledge – until all the collectable data are in and all the possibilities are estimated.
Secondly, a decision of the kind we are interested in for our iGEM project is not the only possible application of the construction of the posterior. For instance the construction of the posterior might be a simple intermediary step in a more complex algorithm - the Bayesian approach is highly modular.
Imagine we conduct a small experiment in order to learn something about a phenomenon – say the growth rate of a colony. If the experiment is very thorough and very well-controlled we can reasonably hope to obtain a reliable estimate of the growth rate. Else, it is dangerous to trust the data too much. On the other hand, it would be wasteful to get rid of the data altogether. But we can still use them as a source of prior knowledge. In practice, the posterior of the growth rate is then used as a prior in the new, more complex process.
The **lesson of the story** therefore is:

- The posterior is what matters
- Be flexible
- Do not make 'hard' choices unless and until you are forced to

Bibliographical Research on Bayesian Parameter Estimation

Now for a problem like ours we need to reach a decision at some point!
As you must surely have realised there are several ways to estimate the parameters from their posterior distribution. This may seem confusing but there is a very good reason for this: there are several ways to estimate the parameters because there are several ways to ask the question 'what is the parameter?'.
In decision theory the outcome to the decision process is characterised by a series of desirable properties – the properties of interest being encoded into a utility/loss function.

- What is a utility/loss function?
- What is a Bayesian estimator?
- What is **MAP**? What are the pros and cons of the method?

Note: The use of a Bayesian estimator is the approach that I encourage you to adopt, but if you find another -justifiable- way to ask the question 'what is the parameter?' to the posterior and you can get an answer , by all means try!

A More Complex Example of Bayesian Parameter Estimation

The exponential distribution only had one parameter to estimate. With several parameters things get a it more complicated.
Now let us deal with a slightly more complex example: the Gaussian distribution. Let us call m and σ its parameters. We assign to σ a prior distribution $P(\sigma)=1/\sigma$ and to m a uniform distribution $P(m)=1$. Again the normalisation constants are overlooked

- Write the general expression for likelihood of the sample vector X
- Write the posterior of (m,σ) - Apply MAP to the posterior of (m,σ)
- Draw 10, 50, 100 samples $X=(x_1,\ldots,x_N)$ from a Gaussian of mean 10, deviation 5. For these samples plot the posterior of (m,σ).
- Calculate its maximum, its expected value and its Hessian

So far, so good: we have simply gone from one dimension to two. However, there is another possibility: we can simply ring ourselves back to one dimension for each of the parameter of the model through marginalisation of the other parameters.

- What is marginalisation?
- Marginalise σ to get the posterior of m
- For the samples you have drawn, plot the Posterior of m.
- And estimate its maximum, mean and standard deviation
- Do the same with the posterior of σ

Bayesian Model Comparison

The real power of the Bayesian approach does not lie in the way it estimates parameters – it is often close the standard way- but in the systematic way it performs higher-level operations such as model comparison.
A beautiful framework was developed by **D.McKay** - See McKay's book ***Chapters 27*** and ***28***- that explains ,among other things, one of the oldest principles of science: Occam's razor

- What is Occam's razor?
- How can the evidence help you rate models and Why?

To see how this works, let us go back to the example of the exponential distribution. Use for the data D the 50 samples you have drawn

- Estimate the evidence of your data D – use both a Laplace approximation and the direct computational approach
- Now consider a uniform distribution between 0 and 10. What is the evidence of the data for such a model?
- Which one is the better model? (please prove it is the exponential…)

The estimation of the evidence can be made much more simpler (to the detriment of precision of course) by considering the degenerate case and fix the parameters to their estimated value

- What is the Evidence?
- Is the exponential model still the better model?

*We now have a way to match models to data and rate the models versus these data…**We have everything we need!***

Multilayered Bayesian Data Analysis

We are now going to combine all the bits that you have developed into a bi-layered Bayesian data analysis. At the first layer, the routine will associate to a model of your choice the best parameters. At the second layer it will determine among a library of model which one is best supported by the data.

Generation of Data

First you will need some data

- Generate motility data according to the model of your choice
- Split them into velocity, run-time, rotation-time and angle
- Do the same with another – unrealistic model.

First Layer

We are only going to analyse the velocity during the run-phase.

- Identify several plausible models for the velocity during the run-phase.
- Estimate the parameter(s) for each model
- Estimate the reliability of your predictions

Second Layer

For the velocity during the run-phase:

- Compute the Evidence for each candidate model
- Determine the best supported model

Design of Experiment

While we are at it, let us have a look at the design of experiment.
We want to estimate the parameters with a 10% precision.

- What does it mean in terms of posterior?
- For some of your candidate models, how many samples do you need to achieve such precision? **Hint:** Use synthetic data.

Unfortunately we cannot be sure of the quality of the data. Let us have a look at how bad data degrade the quality of the first layer.

- Mix the data generated by your realistic model with data generated with your unrealistic model (use a ratio good/bad of 10 , 5 then 2)
- Under the assumption of the correct model, estimate the parameters of the model. Compute the corresponding Evidence.
- What are your conclusions?

At the End of this Tutorial you should…

- Understand the fundamentals of **ML** parameter estimation
- Be familiar with the basics of the Bayesian approach
- Be able of calculating the posterior of the parameters of a distribution
- Be able to use the posterior to estimate the parameters of a distribution
- Be able to use the posterior to quantify the reliability of your predictions
- Be able to compare how well a model fits your data